Association for Information Systems
AIS Electronic Library (AISeL)

ICEB 2018 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-6-2018

Credit Fraud Risk Detection Based on XGBoost-LR Hybrid Model

Maoguang Wang
Central University of Finance and Economics, China, mgwangtiger@163.com

Jiayu Yu
Central University of Finance and Economics, China, jiayuy1212@163.com

Zijian Ji
Central University of Finance and Economics, China, 18146573209@163.com

Wang, M., Yu, J., Ji, Z. (2018). Credit fraud risk detection based on XGBoost-LR hybrid model. In *Proceedings of The 18th International Conference on Electronic Business* (pp. 336-343). ICEB, Guilin, China, December 2-6.

Credit Fraud Risk Detection Based on XGBoost-LR Hybrid Model

(Full Paper)

Maoguang Wang, Central University of Finance and Economics, China, mgwangtiger@163.com
Jiayu Yu*, Central University of Finance and Economics, China, jiayuy1212@163.com
Zijian Ji, Central University of Finance and Economics, China, 18146573209@163.com

ABSTRACT

For a long time, the credit business has been the main business of banks and financial institutions. With the rapid growth of business scale, how to use models to detect fraud risk quickly and automatically is a hot research direction. Logistic regression has become the most widely used risk assessment model in the industry due to its good robustness and strong interpretability, but it relies on differentiated features and feature combinations. XGBoost is a powerful and convenient algorithm for feature transformation. Therefore, in this paper, XGBOOST can be used to effectively perform the advantages of feature combination, and a XGBoost-LR hybrid model is constructed. Firstly, use the data to train a XGBoost model, then give the samples in the training data to the XGBoost model to get the leaves nodes of the sample, and then use the leaves nodes after one-hot encoding as a feature to train an LR model. Using the German credit data set published by UCI to verify this model and compare AUC values with other models. The results show that this hybrid model can effectively improve the accuracy of model prediction and has a good application value.

Keywords: Credit risk, XGBoost, logistic regression.

*Corresponding author

INTRODUCTION

With the development of the economy and the change in the concept of personal consumption, the credit business has gradually entered people's daily lives. According to the Central Bank's financial statistics report, RMB loans for the whole year of 2017 increased by RMB 1.353 billion. Credit business is not only an important business of banks, but also a business that many financial institutions rely on to survive and develop themselves. The credit business is actually a kind of credit granting. The credit agency investigates the borrower's ability to repay and the willingness to repay the loan, and evaluates it to determine whether grant the loan and the duration and amount of the loan are to be determined. In credit risk, the most important is fraud risk that fraudster has no willingness to repay when applying for a loan. In order to cope with this risk, bank managers must find effective solutions in order to distinguish bad ones from applicants. Therefore, with the rapid expansion of the scale of credit business, fraud risk detection of lenders has become a very important task for banks and financial lending institutions.

Fraud risk detection allows the credit industry to benefit from improved cash flow, guaranteed credit collection, reduced potential risks, and better management decisions(Lee Tian-Shyug *et al.*, 2002). Obviously, for a large-scale lending business, the accuracy of the scoring model only increases by 1%, and it will bring significant benefits to credit institutions. Therefore, the research on the credit fraud detection model has practical significance.

RELATED WORK

The credit fraud detection model is actually a classification issue. It classifies lenders into two categories: "good credit" that is capable of fulfilling economic obligations, repaying loan interest interest, and "bad credit" that is rejected grant loans because of high probability of default (Bastos J, 2007).Traditional credit assessment mainly depends on professionals with rich experience in risk management to manually review the basic information and financial information of lenders. However, with the rapid increase in business volume, it is unrealistic to rely on traditional manual methods. The increase in the number of loan applicants has promoted the development of technologies that use quantitative models for assessment, making credit approval procedures automated and efficient, and capable of supervising borrowings.

Durand (1941) first analyzed the credit risk and for the first time built a consumer credit evaluation model using statistical discriminant analysis. The results show that the use of quantitative methods can achieve better predictive ability. Subsequently, Altman used a multivariate discriminant analysis method to create a classic Z-Score method to derive the probability of an enterprise's risk. This model has a good applicability in predicting the financial status of the company.

Traditional linear discriminant methods require data to satisfy a certain distribution and subject to strict assumptions. However, most of the data in the real world do not satisfy the above conditions, and there is a certain degree of collinear relationship between variables. Logistic regression is a credit scoring method that can effectively deal with binary classification problems. Due to its strong interpretability, it is widely used (Zhu Xiaoming & Liu Zhiguo, 2007). However, Harrell and Lee (1985) found that although the assumptions of the linear discriminant method are satisfied, Logistic regression can be as effective and accurate as LDA. LDA (Latent Dirichl*et al.*location) is a document theme generation model, also known as a three-layer Bayesian probability model, which contains three-layer structure of words, topics and documents. The so-called generation model, that

is, we believe that each word in an article is obtained through a process of "choosing a topic with a certain probability and selecting a certain word from the topic with a certain probability". The document to topic follows a polynomial distribution, and the subject to the word follows a polynomial distribution. Qin Wanshun and Shi Qingyan (2003) selected credit card data from a branch of a large commercial bank and constructed a personal credit scoring model based on Logistic regression to better distinguish “good customers” and “bad customers”.

With the popularity of artificial intelligence, a large number of machine learning methods are gradually being applied in the credit domain. David West (2000) studied five neural network architectures: MLP、MOE、RBF、LVQ and fuzzy adaptive resonance theory neural network (FAR). Among the five structures, hybrid expert systems and radial basis neural networks have good performance. At the same time, this study also shows that logistic regression is a good alternative for neural models. Logistic regression is slightly more accurate than the average neural network model which includes some poor neural network training iterations.

However, a single model has its own advantages and disadvantages in terms of robustness, accuracy, and interpretability. Combining each model can combine the advantages of different models, and it is beneficial to enhance the model's generalization ability and enhance robustness.

Peng Runze (2017) uses SVM, RF, ANN, and GBDT as primary classifiers, and the secondary classifiers use logistic regression to build a two-level integration model. He found that the hybrid model is superior to a single machine learning model.The basic model of the Support Vector Machine (SVM) is to find the best separation hyperplane on the feature space so that the positive and negative sample intervals on the training set are the largest. SVM is a supervised learning algorithm used to solve the two-class problem. SVM can also be used to solve nonlinear problems after the introduction of the kernel method. Random Forest (RF) is an integrated learning model based on decision tree. It contains multiple decision trees trained by Bagging integrated learning technology. When inputting samples to be classified, the final classification result is determined by the vote of the output of a single decision tree. Random forest has good tolerance to noise and outliers, and has good scalability and parallelism for high-dimensional data classification problems. In addition, random forest is data-driven nonparametric classification method that requires only prior knowledge by training the classification rules for the learning of a given sample.GBDT is an algorithm that classifies or regresses data by using an additive model and continuously reducing the residuals produced by the training process.GBDT generates a weak classifier for each iteration through multiple iterations, and each classifier trains on the basis of the residuals of the previous classifier. The requirements for weak classifiers are generally simple enough and are low variance and high bias. Qin Wanshun and Shi Qingyan (2003) trained neural network models and combined the obtained scoring results as one of the explanatory variables with the remaining characteristic variables, then applying together to the Logistic regression model. However, due to the high correlation between the results obtained by the neural network and the dependent variables, the remaining features become unimportant, and multicollinearity problems are easily generated. Chen Qiwei *et al.* (2018) used XGBoost as a base classifier to construct a credit scoring model using data space, feature space, and algorithm parameter strategies, and then used a bagging method to learn a good integration model. Cai Wenxu, Luo Yonghao and Zhang Guanxiang (2017) combined leaf nodes that the GBDT generated with original features to form new features and input them into the LR model. Experiments have verified that the prediction effect of the model is far better than that of a single model.Logistic regression (LR)measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.

Logistic regression has become one of the most commonly used models in the industry because of its simplicity, strong interpretability and good stability. However, the LR algorithm is a linear model and cannot capture nonlinear relationship. It requires a large number of feature transformations to find feature combinations. In 2014, Facebook (2014) introduced the solution through GBDT+ LR to solve feature engineering. The experimental results confirmed that GBDT is a powerful and very convenient method to implement feature nonlinearity and tuple conversion. The XGBoost that evolves and optimizes on the basis of GBDT can also obtain the importance of each feature in the model at the same time. It can be used as a criterion for determining the importance of features and can be used for simple and effective feature conversion and combination of features.

Therefore, in order to solve the shortcomings of the lack of accurate characterization of the nonlinear relationship by the logistic regression, we can use XGBoost to characterize the features and fuse the two models. The leaf nodes produced by XGBoost are feature combinations that describe the original features. By adding nonlinear expressions, the expression ability of the logistic regression model is enhanced, making the model more effective. To the best of our knowledge, there was no study on credit risk assessment by using this model fusion strategy, this paper fills in such a literature gap by fusing LR and XGBoost. In the personal credit assessment, this method makes full use of the advantages of LR and XGBoost, and can solves the inadequacies of a single algorithm, providing a new research idea.

MODEL STRATEGY AND DESIGN

From GBDT To XGBoost

GBDT(Gradient Boosting Decision Tree) was first proposed by Friedman(2002). It uses a serial approach to learn the weak classifier CART regression tree to reduce the bias and improve the accuracy of the prediction.

There is a training sample set:

$$D = \{(x_1, y_1), (x_2, y_2) \dots (x_j, y_j) \dots (x_m, y_m)\}$$

Here m is the number of training samples, x and y represent input samples and sample markers, respectively, and D is a set of training data and corresponding sample markers.

Gradient boosting is an additive model. The final result is the addition of the results of multiple decision tree trees. The additive model can be expressed as:

$$f_n(x) = \sum_{i=1}^{n} T(x_i; \theta_i) = \sum_{i=1}^{n} \beta_i h(x; \alpha_i) = f_{n-1}(x) + \beta_n h(x; \alpha_n) \quad (1)$$

Here $T(x_i; \theta_i)$ represents the ith decision tree and θ_i is the parameter of the ith decision tree. Each decision tree $T(x_i; \theta_i)$ can be represented by $\beta_i h(x; \alpha_i)$, where β_i refers to the coefficient of each decision tree and α_i refers to the parameters of the model.
The loss function can be expressed as:

$$L\big(y, f_n(x)\big) = L(y, f_{n-1}(x) + \beta_n \sum_{k=1}^{q} j_k I(x \in R_k) \quad (2)$$

Here R_k denotes the kth leaf node and j_k denotes the value of the R_k leaf node. When x belongs to the value of the leaf node where R_k is located, the predictive value of this CART regression tree $h(x; \alpha_i)$ for x is j_k.

The idea of GBDT optimization parameters is Gradient, but it uses the idea of gradient descent method in the function space. Each iteration is to fit the residual, and use the loss function in the negative gradient value of the current model to fit a weak CART regression tree.

GBDT needs to be iterated many times to make the model have better accuracy. Chen Tianqi and Guestrin C(2016) have optimized this and proposed the XGBoost model. This model can be accelerated by using CPU multi-threaded parallel processing, and it performs great generalization ability and robustness.

Because the CART regression tree is easier to overfit, XGBoost adds a regularization term to the loss function:

$$\Omega\big(f(x)\big) = \gamma T + \frac{1}{2}\lambda ||w||_2^2 \quad (3)$$

Here T represents the number of leaf nodes, and $||w||_2^2$ represents the L2 regularization of leaf nodes. It can control the complexity of the model by controlling the value and number of leaf nodes. Optimizing γ is equivalent to pruning the tree.

In the feature of splitting selection, each time the attribute is scored using the following information gain formula, then the attribute value with the largest Gain is selected to split the existing leaf nodes.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \quad (4)$$

Here G_L, G_R are the first derivatives of the left and right subtrees respectively, and H_L and H_R are the second derivatives of the left and right subtrees, respectively. λ and γ are the coefficients of the L2 regularization of the corresponding value of the leaf node and the number of leaf nodes in the regularization term.

In addition, XGBoost uses a second-order Taylor expansion for the loss function. In the optimization process, not only the first-order derivative but also the second-order derivative is considered.

Model Design

Because of its good robustness and interpretability, logistic regression is very suitable for credit assessment. However, it is difficult to capture non-linear information, and careful engineering of feature engineering is required to ensure the accuracy of the prediction. In order to be able to use XGBoost for feature combination features, two models are merged: XGBoost and Logistic regression. Using data to train a XGBoost model, the samples in the training data are input to the XGBoost model to get the sample's leaf nodes, and the leaf nodes are used as features to train an LR model.The figure 1 is framework of based on XGBoost-LR ensemble model for personal credit risk assessment.

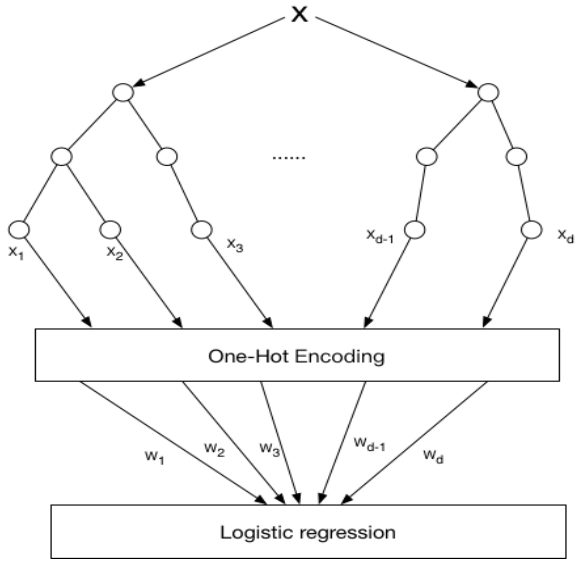

Figure 1: Framework of based on XGBoost-LR ensemble model for personal credit risk assessment.

The specific steps are briefly described as follows:

Step1: Divide the Data Set
Randomly divide the initial sample data into disjoint two parts: the training set and the test set. 80% of the samples were used as training subsets, and 20% of the samples were used as test subsets.

Step2: Data Preprocessing
the attributes in the system have different magnitude, using the data directly affects the results. Subsequently we need to compare the generalization ability of different models, and SVM is a distance-based algorithm, susceptible to the gap between the magnitude of the data. Therefore, the data needs to be normalized. This paper has adopted the maximum and minimum normalization method for the original data.

Step3: Build the XGBoost Model
For each sample, input to the CART regression tree, the value of the falling leaf node as the default probability, and then use the loss function in the negative gradient value of the current model as the residual to fit the next regression tree until finishing training.

Step4: Build XGBoost-LR model
After the sample is input into the XGBoost model, the value of the falling leaf node is coded as 1, the value of the non-falling leaf node is coded as 0, and the output of each leaf node is combined into a feature vector. The resulting vectors are used as the variables of Logistic regression to train the final hybrid model.

EXPERIMENTS

Data Preprocessing

As the current domestic personal credit data is difficult to obtain and confidentiality is strong, most studies mainly focus on the assessment of corporate credit risk, while less research on individual credit risk. This paper uses UCI public the German credit data set (1994) for model validation. The sample size in the data set is 1000, and the ratio of "good credit" customers to "bad credit" customers is 7/3.

There are 20 variables in the original data set, including 7 numerical variables and 13 qualitative variables. Ordered qualitative variables can be directly divided into numbers. Unordered qualitative variables need to be transformed into dummy variables using one-hot encoding. In the end, there are 24 variables, 1 dependent variable, and dependent variable = 1 means “good credit”, dependent variable = -1 means "bad credit".

Model Training

Fraud risk detection is a binary problem, so in the XGBoost model, the negative binomial log likelihood is used as a loss function:

$$L\big(y, f(x)\big) = ln\big(1 + exp\big(-yf(x)\big)\big), \quad y \in \{-1, 1\} \tag{5}$$

As mentioned before, XGBoost belongs to a classifier with a large number of hyper-parameters. The value of the parameter is crucial. Therefore, it is necessary to carefully select the parameter value. However, so far, there are only some heuristic methods,

there is no theoretical method to guide the choice of parameters. In this paper, the grid search method is used to select the parameters. The range of values for the hyper-parameters is shown in Table 1.

Table 1. Range of XGBoost Parameters in Grid Search

Parameter	Grid Search
Number of estimators	(20, 100, 10)
Learning rate	(0.01, 0.2, 0.01)
max_depth	(3, 10, 2）
min_child_weight	(1, 6, 2)
Column subsample ratio	(0.5, 1, 0.1)
Subsample ratio	(0.5, 1, 0.1)
Gamma	(0, 0.1, 0, 01)
lambda	(0.1, 0.8, 0.1)

Grid search uses 10-fold cross validation to perform parameter optimization. The training data is randomly divided into 10 disjoint subsets. Each subset is used as a validation set, and the remaining 9 subsets are used as training sets. After parameter optimization, the structure of the model was determined to be 40 CART regression trees. The learning rate of each tree was 0.01, the max_depth of the tree was 7, the Column subsample ratio was 0.6, and the Subsample ratio was 0.7. In order to prevent XGBoost's powerful fitting ability from overfitting, the L2 regularization penalty term was added. The lambda parameter optimization result was 0.3.

In terms of criteria, the type 1 error rate, type 2 error rate, and AUC on the test set are used as performance evaluation criteria for the model.The type 1 error rate indicates that there is a personal credit risk misjudged as no credit risk, and the type 2 error rate indicates that no credit risk is predicted to have a personal credit risk. In the area of credit risk assessment, there will be higher costs for misidentifying individual credit risk as having no credit risk, so the Type 1 error rate needs to be focused.

Model Results

In order to verify the validity of the proposed XGBoost-LR model, the paper compares the results of the five models: logistic regression classifiers, random forests, XGBoost, SVM, and naive Bayes algorithm using the original features. And compared with the model proposed by Cai Wenxu, Luo Yonghao and Zhang Guanxiang (2017) (marked as GLR), both using the same data training set and test set for training. All compared single models apply the default values of the model in the scikit-learn package.

Table 2. Performance comparisons for different classification and ensemble models.

Model	type 1 error rate (%)	type 2 error rate (%)	AUC(%)
SVM	7.75	81.03	74.53
Naive Bayes	25.35	37.93	71.6
Random Forest	13.38	68.97	71.98
XGBoost	8.45	65.52	77.34
LR	7.75	81.03	73.81
GLR	3.57	70.00	82.68
XGBoost-LR	**3.57**	**65.00**	**83.21**

From Table 2, it can be seen that in the five single models using the original features, the AUC value of XGBoost is the highest, which is 77.34%. Although Naive Bayes' error rate of Type 2 is lower than XGBoost, its Type 1 error rate is higher than XGBoost. However, the cost of type 1 error rate is much higher than that of type 2 error rate. In practice, it is necessary to control type 1 error rate. Therefore, overall, XGBoost outperform naive Bayes. The performance of SVM and LR is not much different from that of XGBoost, indicating that Logistic regression model is a competitive classification model.

For type 1 error rate and AUC, the two mixed models GLR and XGBoost-LR perform best, indicating that the combination of features can effectively improve the generalization ability of LR and fit the true distribution in the higher dimensional space, making the prediction more acurrate. However, the GLR's AUC value is lower than that of XGBoost-LR, which also confirms that Shi Qingyan (2005)stated that the results obtained from the previous model have a high correlation with the dependent variables, combined with the remaining feature variables, and applied together to Logistic regression model is prone to multicollinearity problems.

Therefore, the leaf nodes obtained by XGBoost as variables of Logistic regression after one-hot encoding has the validity and feasibility.

Paired T-Test

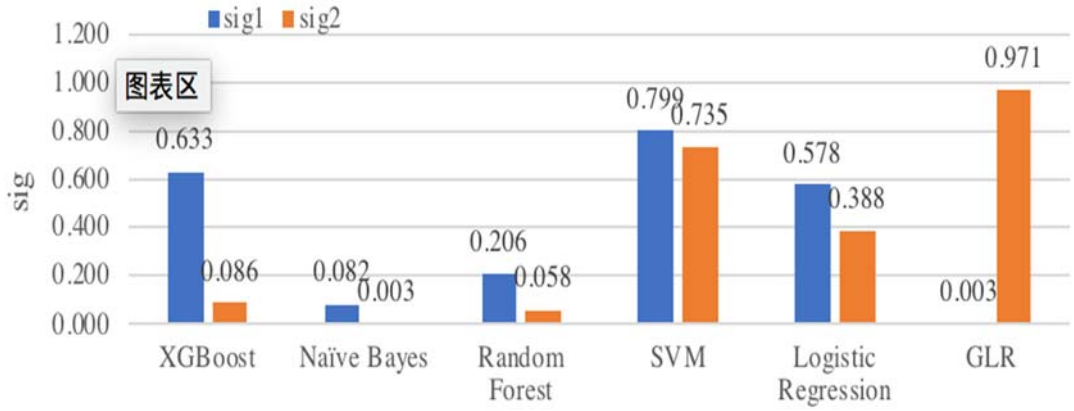

Figure 2: Sigs of paired-t-test of xgboost with other models with α=0.05

In order to make the experimental results more reliable, the paired-t-tests were conducted respectively between XGBOOST-LR hybrid model and others models on the experimental results. The results are shown in Figure 2. Among them, sig1 represents the correlation of two sets of experimental data. And most sig1 values are greater than 0.05, indicating that there is no significant correlation between the two models. While the sig1 of GLR is less than 0.05, which is also consistent with the similarity of the two models. The sig2 value we are more concerned with is the key t-test result. If it is less than 0.05, there is a significant difference between the two groups. As can be seen from the above figure, the sig2 value of Bayes model is less than 0.05, indicating that there is a large difference between XGBOOST-LR hybrid model and Bayes. While the differences between GLR, SVM, LR and XGBOOST-LR are not significant. Considering that the mean error rate of XGBOOST-LR is the smallest one in the experiment, the test results also show that the XGBOOST-LR hybrid model is better than other models.

Interpretability

Interpretability is a very important part of fraud risk detection. Considering interpretability, the entry rules can be made, such as: address blacklist, phone blacklist, the number of loans is greater than a certain number to reject the applicant. So before calling the model, you can reject the application that is likely to be a fraudster, which reduce the cost of calling the model. Second, for banking financial institutions, they need to know which variables are more important in credit risk assessment and can be given more attention.The interpretability of XGBoost model mainly depends on two aspects: feature importance score and decision rules (Xia Yufei *et al.*, 2017).

After training in the XGBoost model, the importance of features is shown in Fig. 3. The larger the F value, the more important the features. The values of F3 (Purpose), F9 (Other debtors / guarantors), and F1 (Duration in month) are relatively large, which is consistent with intuition. The longer the loan period, the higher the likelihood of default. Therefore, in the credit risk detection, more attention needs to be paid to these three variables.

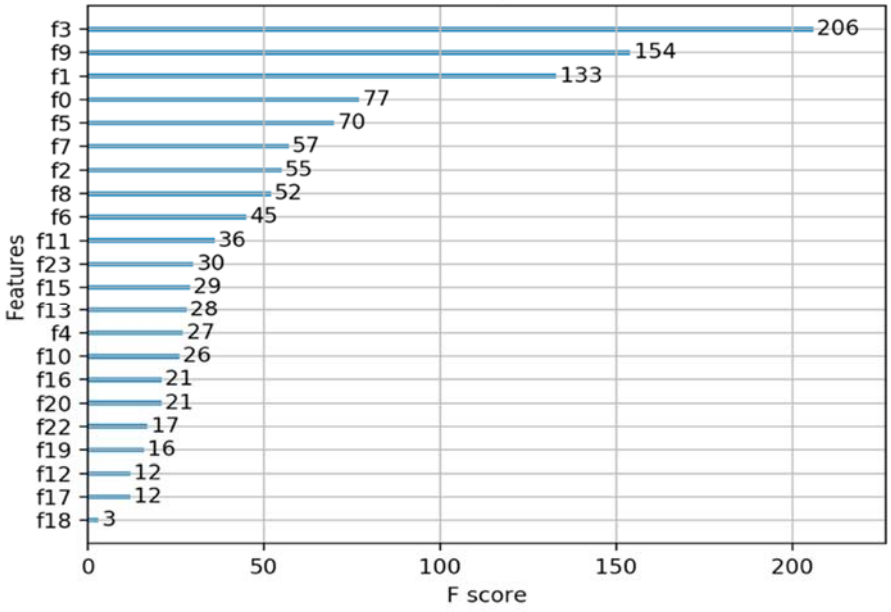

Figure 3: Feature importance scores.

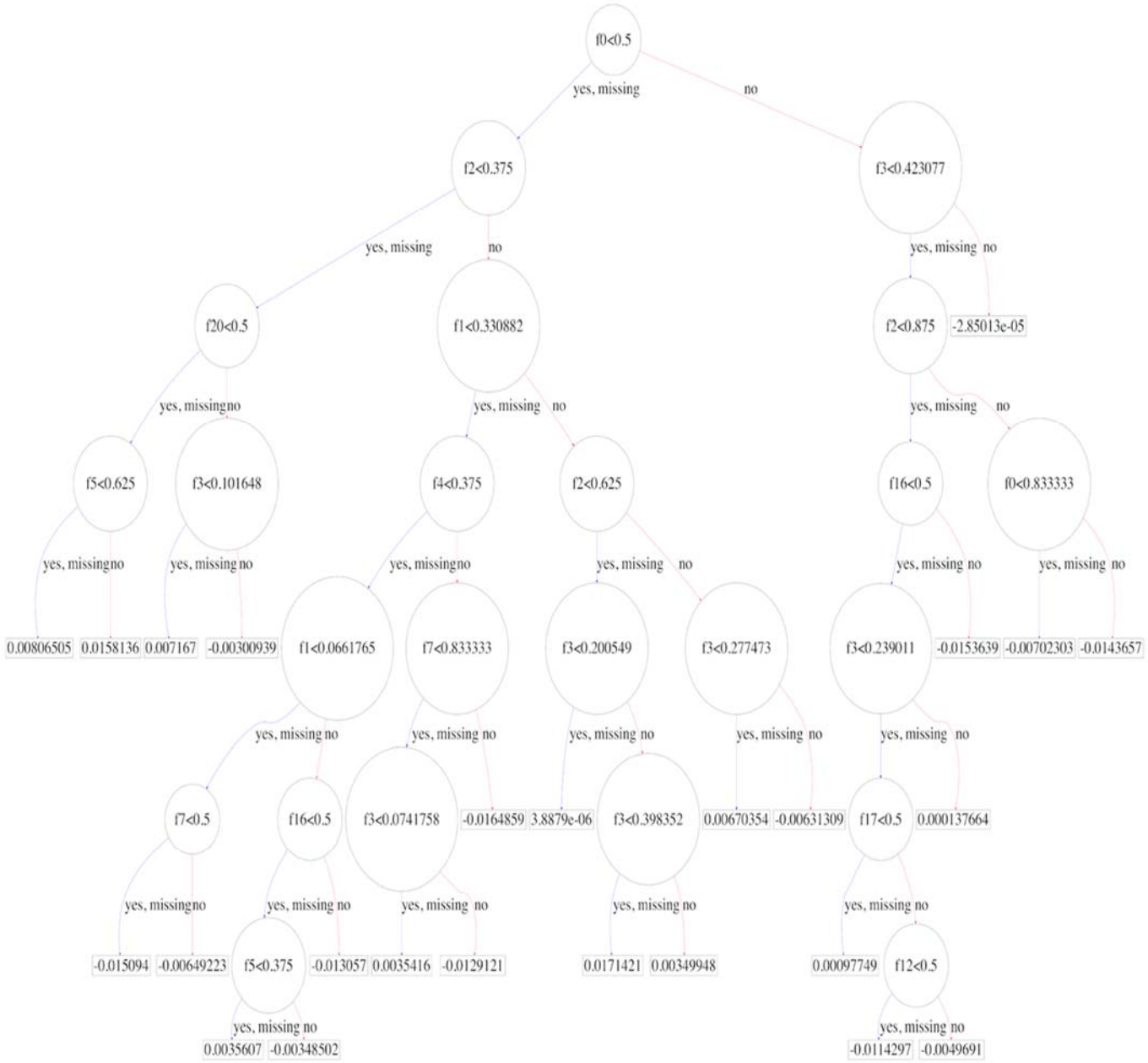

Figure 4: No.40 CART tree of XGBoost-based credit scoring model

The final XGBoost model was obtained by adding 40 CART regression trees. The 40th CART regression tree is shown in Fig. 4. Decision diagrams provide managers with clear and comprehensive decision rules. Applicants are divided into corresponding leaf nodes through a series of decision rules to decide whether applicants is the fraudster. This is a useful tool in the case of real credit scores.

CONCLUSION

With the expansion of the scale of the credit industry, personal fraud risk detection has become increasingly important. A model with a high accuracy rate can bring back significant losses to banks and financial institutions. XGBoost can be used to mine effective features and feature combinations. XGBoost is different from the random forest which is parallel combination strategy. It adopts a serial method and uses the loss function in the value of the current model's negative gradient approximation to fit the regression tree. The result verifies that XGBoost has better generalization performance than the random forest. The LR model is characterized by its simple operation, good robustness, and strong interpretability. However, its accuracy depends on the variables of the input. Then the feature engineering often depends on manual operations before. This requires experience and is not easy to have good results.

This paper uses the advantages and characteristics of the XGBoost and LR models for model fusion. Each leaf node of XGBoost is a combination of some features. By using all leaf nodes as features of the LR model, effective features and feature combinations are mined. The results also show that this hybrid model can improve the performance of a single model and have better generalization capabilities.

The future direction of work will focus on considering how to deal effectively with sample imbalances in this hybrid model., and using domestic credit data to better construct an XGBoost-LR hybrid model suitable for real credit business in China.

ACKNOWLEDGMENT

This work is supported partially by NSFC under the Agreement No. 61073020 and the Project of CUFE under the Agreement No. 020674115002, 020676114004, 020676115005.

REFERENCES

[1] Bastos J.Credit scoring with boosted decision trees[OL]. (2007-05-27)[2008-04-02]. http://mpra.ub.uni-muenchen.de/8156/.
[2] Cai Wenxu, Luo Yonghao &Zhang Guanxiang (2017).Personal Credit Risk Assessment Model Based on Integration of GBDT and Logistic Regression and Empirical Analysis. Modernization of Management:(2):1-4.
[3] Chen Qiwei, Wang Wei, Ma Di, *et al.* (2018) Class-imbalance credit scoring using Ext-GBDT ensemble[J]. Application Research of Computers: 35(2): 421-427(in Chinese)
[4] Chen Tianqi, Guestrin C(2016) XGBoost: a scalable tree boosting system [C]//Proc of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco: ACM.
[5] David W(2000). Neural network credit scoring models. Computers and Operations Research: 27 (11) :1131-1152
[6] Durand D(1941)Risk elements in consumer instalment financing. New York: National Bureau of Economic Research: 189-201
[7] Friedman J H(2002) Stochastic gradient boosting. Computational Statistics & Data Analysis: 38(4): 367-378
GBDT and Logistic Regression and Empirical Analysis. Modernization of Management:(2):1-4.
[8] Harrell F E, Lee K L (1985) A comparison of the discrimination of discriminant. analysis and logistic regression under multivariate normality. Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences :333-343
[9] He X, Pan J, Jin O, *et al.* (2014) Practical Lessons from Predicting Clicks on Ads at Facebook[C] //Proc of 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York :ACM
[10] Lee Tian-Shyug, Chiu Chih-Chou, Lu Chi-Jie, *et al.* (2002) Credit scoring using the hybrid neural discriminant technique. Expert Systems with Applications 23 (3) :245-254
[11] Peng Runze(2017). Personal Credit Assessment Model Based on Stacking Ensemble Learning Algorithm. Statistics and Application: 6(4):47-54 (in Chinese)
[12] Qin Wanshun. Shi Qingyan(2003) Personal Credit Grading Model Basing On Logistic Regression. Quantitative Economics in the 21st Century: 4 (in Chinese)
[13] Shi Qingyan(2005). The Research of A Mixed Two-Phase Scoring Model of Personal Credibility Based on Neural Network-Logistic Regression[J]. STATISTICAL RESEARCH :22 (5) :45-49(in Chinese)
[14] UCI(1994). Statlog (German credit data) data set. https://archive.ics. uci.edu/ml/machine-learning-databases/statlog /german.
[15] Xia Yufei, Liu Chuanzhe, Li YuYing, *et al.* (2017) A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring[J]. Expert Systems With Applications:78: 225-241
[16] Zhu Xiaoming, Liu Zhiguo (2007) Credit Score Model Review. STATISTICS AND DECISION 2:103-105 (in Chines)